DOCUMENT RESUME

ED 441 339 FL 026 260

AUTHOR Pollard, John D. E.
TITLE The Influence of Assessor Training on Rater-as-Interlocutor Behaviour during a Computer-Resourced Oral Proficiency Interview-Cum-Discussion (OPI/D) Known as the "Five Star Test." Paper II.
PUB DATE 1998-01-19
NOTE 10p.; Paper presented at the Current Trends in English Language Testing (CTELT) Colloquium (Al-Ain, United Arab Emirates, 1998). For Paper I, see FL 026 259.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *English (Second Language); Evaluation Methods; Foreign Countries; Interviews; *Language Proficiency; *Language Tests; *Oral Language; Second Language Instruction; Second Language Learning; Test Construction; *Test Validity; Testing
IDENTIFIERS *Oral Proficiency Testing; *Rater Effects; Saudi Arabia; United Arab Emirates

ABSTRACT

This paper considers the influence of the values and the discourse behaviors of the native-speaking assessor as interlocutor. These variables in the assessments of oral proficiency interrelate in complex ways with features of test design such as tasks, prompts, topics, guidelines, and assessor training. This paper reviews some previous studies, and looks at measures taken in the design of the "Five Star Test" to improve discourse and interactive consistency. An exploratory test suggests that assessor training might be a more beneficial area of attention for test developers than recent research indicates, especially when coupled with innovative features of test design. Previous research has shown that oral proficiency interviews could and should be made more like real conversations. This paper aims to take steps in that direction by incorporating innovative features into proficiency instrument design relating to aspects of test design, theme and topic, task design, and assessor/interlocutor training. It is concluded that, because the assessor "gaze" in signaling commitment and participation to the student being assessed has measurable effects on student performance in the second language, developmental investment in assessor training and awareness-raising would contribute to both test validity and reliability. (Contains 27 references.) (KFT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 339

**Paper II : The influence of assessor training on rater-as-interlocutor behaviour during a computer-resourced Oral Proficiency Interview-cum-Discussion (OPI/D) known as the 'Five Star Test'.**

John D.E. Pollard

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

John Pollard

BEST COPY AVAILABLE

FL024260

**Paper II : The influence of assessor training on rater-as-interlocutor behaviour during a computer-resourced Oral Proficiency Interview-cum-Discussion (OPI/D) known as the 'Five Star Test'.**

This considers the influences of the values, and the discourse behaviours of the NS Assessor as Interlocutor. These variables in assessments of oral proficiency interrelate in complex ways with features of test design such as tasks, prompts, topics, guidelines, and assessor training. This paper reviews some previous studies, and looks at measures taken in the design of the Five Star Test to improve discourse and interactive consistency. An exploratory study suggests that assessor training might be a more beneficial area of attention for test developers than recent research indicates – particularly when coupled with innovative features of test design.

Assessors of oral proficiency have been shown to carry preconceived, internalised, and perhaps prescriptive, notions of proficiency which operate on their judgements independently of band-descriptors and in spite of guidelines provided by examining and testing bodies. (Ludwig, 1982; van Lier, 1989; Barnwell, 1989; McNamara, 1990; Ross, 1992; Brown, 1993; Chalhoub-Deville, 1995). This may result in a failure to credit, or even a tendency to penalise learners for behaviours that constitute recognised models of second language performance (Canale & Swain, 1980; Lantolf & Frawley, 1988; Bachman, 1990; Alderson, Clapham & Wall, 1995; Bachman, & Palmer, 1982, 1984, 1997).

The problems for test reliability resulting from this are compounded in many OPIs (including 'Five Star') where the assessor and interlocutor are the same person. The Steven Ross study cited above, for example, shows that proficiency ratings vary inversely with the amount of accommodation offered.

In the 1989 article, van Lier expressed the view that OPIs could be made more like real conversations. He did not give detailed indications of how this might be achieved, beyond citing an example of test design where more consideration than usual was given to the themes and topics of the tasks (van Lier 1989: 501-2), and pointing researchers in the direction of Conversation Analysis. Most of the research that has followed this lead has been based on tests that are already long-established. This is inevitable, given the nature of research funding. However, it means that the first part of van Lier's proposition remains unexplored. This paper aims to take a very tentative step towards rectifying this, by basing an enquiry on a test that has been developed with a number of innovative features which are, to the best of my knowledge, currently unavailable to the proficiency instruments of major testing bodies such as ACTFL, IELTS and UCLES. These include aspects of

- Test design – including rating procedures developed with reference to the demands made on the Assessor-cum-Interlocutor.
- Theme and Topic – with reference to the importance placed on local, cultural and personal saliencies.
- Task Design – referring to the creation of 'role-reversals', two-way information gaps, and the establishment of clear and unobtrusive performance criteria.
- Assessor / Interlocutor Training – with an analytical glimpse of what can happen when there is none.

Test Design

One difficulty for OPI Assessor/Interlocutors appears to be their inability to relate knowledge of scale descriptors to actual performance. Band-descriptors are by nature generalisations. Many include expressions like 'when discussing a familiar event', 'in everyday conversation', etc. It is understandaoly difficult to compare real and particular instances of performance with such statements. Applying these descriptors whilst actually being engaged in the complex processes of interaction with the candidate further complicates the event. In the case of some OPIs (e.g. ACTFL OPI, Cambridge CASE) there is yet another requirement – namely, suppressing features of interaction which appear to be quite natural to non-test NS-NNS encounters – such as slowing down one's speech and supplying items that seem to be 'on the tip' of the candidate's tongue, or 'collaborative completion'. (Perret, 1990; Lazaraton, 1996: 154-5). Other tests require the assessor to refer to an interview format and/or evaluation

criterion during the actual process of the OPI. There are recorded instances where this inauthenticates the interaction when compared with non-test exchanges. For example, the candidate 'does some further topic talk on his name, . . . but all that he gets in response . . . are three weak agreement markers' because the interviewer is preoccupied at the time with his interview agenda. (Lazaraton, 1992: 378)

The philosophy underpinning the design of the Five Star test is that both the assessor and the candidate will behave more naturally, if the assessor is relieved of the burdens of monitoring the candidate and consciously deselecting responses, and if vague and obtrusive guideline and evaluation interventions are removed from the event. Support for this can be found in work on Interlanguage (Selinker, 1972; Higgs & Clifford, 1982; see also Underhill, 1987 and Pollard, 1994). This is implemented through design features such as the use of pop-up on-screen text-boxes, which, after the initial assessor training, serve only as reminders of task requirements and evaluation criterion. In actual test use they rarely need to be accessed, and then only briefly, causing minimal distraction of the assessor from his / her engagement with the candidate. The auto-scoring mechanism which multi-functions as a task navigator is equally unobtrusive to the process and inconspicuous to the candidate.

These features can be seen in the first task, depicted below. At the bottom of the task screen, in very low profile, are three evaluation buttons which appear as arrows. To the left of each arrow is an icon (shown below as a square) 'pointing-and-clicking' which reveals pop-up scripted evaluation criteria, as shown in the speech bubbles. As mentioned above, these are used during assessor-training, and only *in extremis* during real tests.

In this way specific rather than global evaluation decisions are made; they are made on a local task-by-task basis rather than cumulatively; and they are made instantly rather than retrospectively. At the end of each individual test algorithm, (consisting of between 12 and 20 tasks) a histogram graph and set of band-descriptors for the cumulative performance across the constructs tested, are automatically generated.

There is no empirical evidence for this claim, other than an intra- and inter-rater reliability study referred to in Paper I – but anecdotal feedback suggests these features help to ease the burden on assessor or rater reliability.

Task.
Each task on the Five Star Test has been designed with a specific criterion which, on the basis of pre-trialling, discriminates performance into one of three broad categories. (Pollard, 1994). The test is algorithmic and adaptive, so that all candidates begin at the same point, and then branch according to performance. The replication throughout the test of task types, supporting graphics, and screen configurations is employed to ease the burden on the assessor, reducing his/her need to access the on-screen pop-up guidance. This and the gradual introduction of multimedia features is also thought to diminish method effect due to lack of familiarity on the part of the candidate. For example sound from the computer is first encountered in simple instances in the early stages of a test pathway where there is a considerable amount of additional support.

One of the consistent criticisms of OPIs has been the lack of symmetry in the discourse they generate, which, at one extreme has been described as being more like an interrogation than a conversation (van Lier, 1989; Young & Milanovic, 1992; Zeungler, 1993; Young, 1996). Through the use audio-recorded of L1 (Arabic) instructions, tasks can be set up where the information for proceeding with the test is given to the candidate rather than the assessor. For example, immediately following the initial task which is based around the registration of the candidate's details, the candidate is instructed (in L1) to gather similar information for on-screen boxes which are designated (First Name, Second Name, Family Name, Nationality, etc) in Arabic. These have to be entered in English by the assessor under the guidance of the candidate, effectively reversing the discourse initiative without requiring the candidate to use the computer. In other instances, task instructions are given in Arabic, and the candidate has to explain them to the assessor. This explanation forms a Speaking and Interaction task in its own right. Both of these are instances of what have been referred to in second language

classroom contexts as two-way 'information gaps' and are designed to authenticate interaction through a real need to communicate. (Doughty & Pica, 1986, Nunan, 1988)

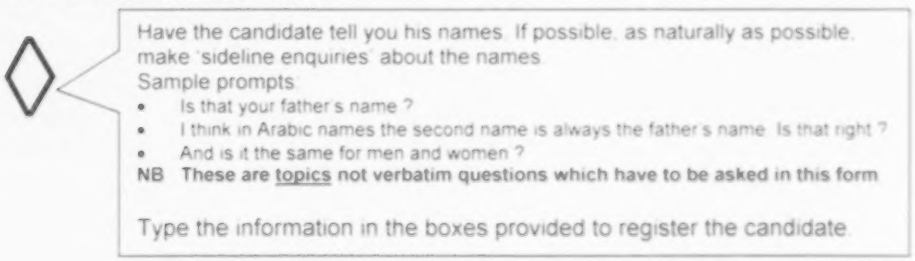
As mentioned above, task-by-task evaluation is one means by which the Five Star process seeks to alleviate the burden on the assessor. The first task, illustrated below, shows how this works. This task combines the registration procedure of obtaining personal details (names, approximate age, birthplace, etc) with specified 'sideline' discussions. As Lazaraton points out, in many OPIs the registration and introduction are assigned as uncredited 'warm-ups', 'losing', for evaluation purposes one of the most authentic phases of the event. (Lazaraton, 1992: 382 )

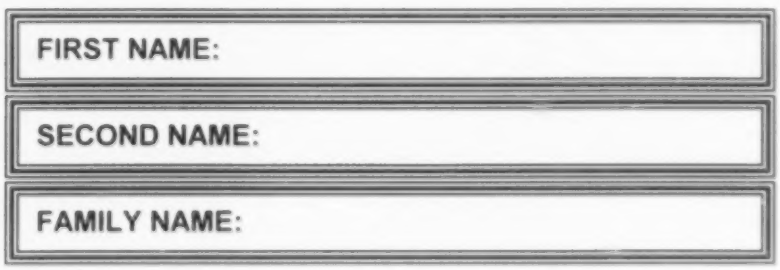
Theme and Topic

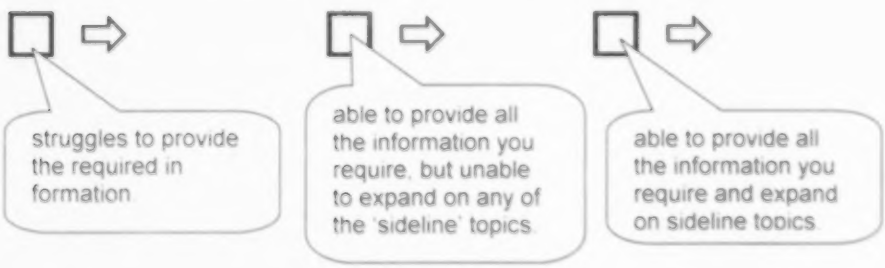
There is research evidence that saliency of topic is a powerful influence on the discourse structure (Woken & Swales, 1989, Milanovic & Young, 1992, Zeungler, 1989, 1993, Young, 1996). In the example below, the 'embedded' task of having 'sideline' discussions is based on a wide range of name-related topics which were trialled for their accessibility to the test population in terms of the language sample they elicited, and the 'naturalness' with which they merged into the 'dominant' task – in this instance, registering biodata. An effort has been made to juxtapose themes of successive tasks which make for natural conversational progressions. For example, the discussion of names mentions locality (since family or 'tribal' names are regional), this leads into a discussion of birthplace and on to schooling experiences. This leads into post-school experiences, and then into travel experiences, etc.

In every task the attempt is to personalise the topic so that the candidate and not the assessor is the 'knower'. It has been shown that when topics are more equally shared between assessor and candidate – as in the case of academic subject specialisms – there can be an affective reaction bound up with self image. (Zeungler, 1989:238)

The following diagram illustrates of some of the developmental measures taken to moderate variables of Test, Task and Topic Design which have been empirically demonstrated to skew the discourse structure of OPIs.


Have the candidate tell you his names. If possible, as naturally as possible, make 'sideline enquiries' about the names.
Sample prompts:
• Is that your father's name ?
• I think in Arabic names the second name is always the father's name. Is that right ?
• And is it the same for men and women ?
NB These are topics not verbatim questions which have to be asked in this form
Type the information in the boxes provided to register the candidate.





struggles to provide the required in formation
able to provide all the information you require, but unable to expand on any of the 'sideline' topics.
able to provide all the information you require and expand on sideline topics.

Assessor / Interlocutor Training

A bigger (and I think more interesting) challenge, however, is defining *Interlocutor Support*, and ensuring that it is consistently offered. This is where Dr Chalhoub-Deville's interest in raters as dimensions of the proficiency both coincides with and departs from my own. Her concern is that raters bring with them their own, internalised evaluation criteria, which operate regardless of guidelines. This concern was also a part of the motivation for van Lier's 'warrant' for an enquiry into Conversational Analysis. The view he expressed in 1989 was that lack of detailed knowledge about the precise things that make *proficient* interactive performance was partly responsible for Assessors falling back on 'criterial linguistic features' (such as microlinguistic accuracies in terms of pronunciation and grammatical formatives) and disregarding instructions to base assessments on successful task completion.

What is the nature of this detailed knowledge, then, that might help us out of the difficulty ? The tendency in research triggered by van Lier's seminal observations has been to identify tokens of contingency (from the conversational analysts) which shape the discourse structure of samples of audio-recorded OPI assessments. (Young & Milanovic, 1992; Young, 1996). Lazaraton (1992) used some video data, but again the focus was mainly on overall discourse structure, and the test in focus had design features which dominated the goal-orientation of the interviewer/assessor and skewed the process towards asymmetric contingency. The 1996 Lazaraton study looks more closely at the turn-by-turn construction with particular attention to interlocutor support, but does so primarily from linguistic perspectives of content and structural formatives. The Conversation Analysts focus much more attention on 'the omnirelevance of action' (Schegloff, 1995) and are more concerned than applied linguists with the paralinguistic features which attach to 'turn-constructional units'. In fact they examine how the linguistic, supralinguistic and paralinguistic interrelate, and recognise *'the sequential relevance to interaction for participants of eye gaze, facial expression, gesture, body deployment, pitch, intonation, vocal stress, orientation to objects in interactional space, laughter, overlap and its resolution, unfinished and suppressed syllables, and silence.'* This reveals the limited scope of extant test research, and of studies which focus only on discourse structure and language – limitations which recent researchers have acknowledged (Young & Milanovic, 1992 422; Young, 1996: 37). Above everything else, it reveals the complex array of rater-as-interlocutor dependent variables.

With that in mind let me share with you some very raw and tentative information – I can't call it data. The background is as follows. Circumstances threw in my way something that I wasn't able to construct in a controlled research environment, though this may be possible in the future.

Managers working in a separate location from where the test was being developed had a requirement for some English proficiency assessments, and asked two new members of our ELT staff to administer the Five Star Test. At this stage the only version available to them was an early prototype version which did not include the assessor guidance and evaluation 'pop-ups'. They were given time to flick through and familiarise themselves with the tasks, but were given no proper training in line with the design.

As a result, no-one had emphasised the importance of maintaining a friendly, non-judgemental mien when conducting the assessment, and it was later established that the teachers were unaware of supportive behaviours which have been studied in OPI, Counselling, and similar contexts. The test developers were powerless in the face of commercial pressures to do anything about this, but the teachers were willing to video some of their tests for later feedback, to 'ensure they were doing it right'. In contrast, and at the same time, other assessors were being trained in Riyadh where the principle development was taking place. The latter had, of course, been fully briefed in the intended approach, and some of their early test performances had been videoed for developmental purposes. Between all the assessors, there were no gender differences and no significant age differences. All had at least three years' experience of living and working in the Gulf. The candidates assessed in the videos were also all-male, all were between the ages of 25-35, and an independent evaluation of the video data estimated that they covered comparable ranges of proficiency.

So, basically, the two pairs of Assessors were conducting themselves according to different briefings and guidelines: the one briefing led them to assume the role of 'tester' in the judgemental sense, and gave no indication that the interactions that took place around specific 'tasks' was a part of the assessment process; the other briefing stressed the importance of friendliness and supportiveness, introduced the assessors to some of the relevant literature, and allowed the pre-set criterion to steer the evaluation.

One of the behaviour variables referred to in the work on Conversation Analysis is 'eye-contact' or, as it is referred to in this field, 'gaze'. This behaviour is generally viewed in interactional terms as a 'display of recipiency' or 'co-participation' . (Atkinson & Heritage, 1984; Goodwin, 1984). On the basis of this, I decided to compare the frequency and duration of Assessor-initiated eye-contact between these assessors.

Method

Ten performances of the same task were identified for each of four assessors – the two who had not been briefed vis-à-vis interlocutor supportiveness, and two from those who had. The tasks were audio-recorded from the video and then transcribed. The transcriptions were printed and then manually 'marked-up' for instances and duration of assessor eye-contact with candidates.

Shared task-boundaries were identified (commencing with a request for the candidate's first name or application number and ending with the onset of the closing move before the assessor moved to the next task.) Importantly, the more discursive parts of this task – those prompted by the 'sideline enquiries' in the pop-up task instructions – were excluded from the analysis. This more 'socially-' than 'functionally-oriented' part of the task was not implemented by the 'untrained' assessors, and its inclusion would have skewed the data.

Data

A table of eight column/cells was compiled for each assessor, all containing, from left to right:

1 - the median of the total task duration, as recommended in Young & Milanovic (1992) and Young (1996) for instances where the data set is small and the range wide. (With such a small data set it could clearly be seen that this was a far better representation of the time each assessor typically spent on the task than the mean would have provided, as well as a better measure for inter-rater comparisons.)
2 - the range, or longest and shortest time dedicated to this task in the samples.
3 - the number of spoken turns (again, using the median).
4 - the range of spoken turns.
5 - the instances of assessor-to-candidate gaze (median).
6 - the range of instances of assessor-to-candidate gaze.
7 - the duration of assessor-to-candidate gaze across all samples, measured in seconds and hundredths of seconds. The means for each assessor across the ten tasks were used, having first omitted rapid 'glances' of less than three seconds duration, as these seemed to be fulfilling a different function than 'engagement and co-participation', and were singularly evident in one assessor.
8 - the range of duration of assessor-to-candidate gaze across all samples, again with the exclusion mentioned in 5.

| Task duration | | Assessor turns per task | | Instances of Assessor gaze | | Duration of Assessor gaze | |
|---|---|---|---|---|---|---|---|
| 1 median | 2 range | 3 median | 4 range | 5 median | 6 range | 7 median | 8 range |
| ASSESSOR 'A' | | | | | | | |
| 01:93:38 | 01:51:04 to 02:48:56 | 30 | 23-35 | 16 | 8-19 | 54:90 | 26:09 to 63:11 |
| ASSESSOR 'B' | | | | | | | |
| 01:42:14 | 01:30:58 to 01:52:45 | 19 | 12-24 | 11 | 2 - 15 | 15:96 | 05:09 to 20:97 |
| ASSESSOR 'C' | | | | | | | |
| 01:50:76 | 01:21:64 to 02:16:46 | 24 | 13 - 31 | 12 | 8 - 15 | 51:54 | 29:60 to 68:11 |
| ASSESSOR 'D' | | | | | | | |
| 01:88:25 | 01:42:87 to 02:03:39 | 24 | 15 - 30 | 12 | 6 - 16 | 32.65 | 05:13 to 40.60 |

The assessors who had been trained in accordance with the philosophy of the test were A and C.

Results

The only data which consistently varies between the trained and untrained assessors is the median figure for duration of assessor-to-candidate gaze, where more 'gaze-time' was given by the trained assessors.

Discussion

Due to the small sample size, these results are only an indication of what might be found through more rigorous enquiry. In addition to seeking more rigorous data and variable controls within the parameters employed here, further potential lies in examining whether, per comparable unit of talk, the putatively more affiliative / less judgemental approach by assessors elicits more interaction, as measured in comprehension checks, requests for clarification, etc. As an ethnographic exercise, groups from the target population could be asked to evaluate the relative styles of the assessors, as well as whether the candidates appear more or less reserved. It is probable that such research might be revealing in terms of cross-cultural paralinguistic behaviours, and might provide specific advice that could be given to trainee assessors.

As referred to in the Assessor A's pattern of eye contact differed from that of B, C and D, in that it included a number of rapid 'glances' which ranged from 00 29 to 03 04 seconds. These appeared to follow questions which were posed without any eye contact (while the assessor was looking at the computer screen). One could speculate that this behaviour emanates from an approach which views the questions as prompts for test performance rather than any desire to engage with the candidate or ask questions to find out information of shared interest

Without investigating it in the necessary detail, it appears that although the questions asked revolved around the task theme (names), they juxtaposed topics within that theme which bore little reference to the candidate's responses. This results in a series of questions and answers more characteristic of interviews in terms of 'asymmetric contingency' than the 'more conversation-like' pattern of reactive contingency. Interestingly, this candidate refers to the process as an interview when commencing the test. He also consistently fails to exploit the opportunities for 'ice-breaking' and bonding – particularly with candidates at the lower end of the range selected, and in one assessment, only made eye-contact twice with the candidate. On one occasion, a candidate felt it necessary to ask if he could know the assessor's name after the assessment had been completed.

On this basis perhaps we can surmise that two of the assessors in this study took on the role of judge but not the responsibilities of being a supportive interlocutor, and that in doing so they initiated and maintained shorter periods of eye-contact with their NNS-candidates. The NS-assessors who supportively engaged their NNS-candidates, and indulged in the collaborative construction of meaning before making an assessment, maintained longer periods of eye-contact with their NNS-candidates. Given the importance of 'gaze' in signalling commitment and participation that has been recorded in NS-NS conversation, it is likely that it plays a part in revealing the co-constructional 'interactive' abilities of L2 learners in oral proficiency assessments. This being the case, developmental investment in assessor training and awareness-raising would contribute to both validity and reliability.

John Pollard, Riyadh, 19/01/1998

References


**Alderson J C, Clapham C and Wall D**. 1995. *Language Test Construction and Evaluation*. Cambridge University Press.

**Atkinson J M and Heritage J (Eds)**. 1984. *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press.

**Bachman L S**. 1990. *Fundamental Considerations in Language testing*. Oxford University Press.

**Bachman L S and Palmer A S**. 1996. *Language Testing in Practice*. Oxford University Press.

**Barnwell D**. 1989. *'Naive' native speakers and judgements of oral proficiency in Spanish*. Language Testing. 6/2. 152-63.

**Canale M and Swain M**. 1980. *Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing*. Applied Linguistics 1/1. 1-47.

**Chalhoub-Deville M**. 1995. *A Contextualised Approach to Describing Oral Language Proficiency*. Language Testing 45/2. 251-281.

**Doughty C and Pica T**. 1986. *'Information-Gap' Tasks: Do They Facilitate Second Language Acquisition ?* TESOL Quarterly. 20/2. 305-325.

**Goodwin C**. 1984. *Notes on Story Structure and the Organisation of Participation*. In Atkinson J M and Heritage J (Eds). 1984. *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press. 223-4.

**Lantolf J P & Frawley W**. 1988. *Understanding the Construct*. Studies in Second Language Acquisition. 10. 181-195.

**Lazaraton A L**. 1992. *The structural organisation of a language interview: a conversation analytic perspective*. System. 20. 373-86.

**Lazaraton A L**. 1996. *Interlocutor support in oral proficiency interviews: the case of CASE*. Language Testing. 13/2. 151-172.

**Ludwig J**. 1984. *Native Speaker Judgements of Second Language Learners' Efforts at Communication: A Review*. Modern Language Journal. 66. 274-283.

**Nunan D**. 1988. *The Learner Centred Curriculum*. Cambridge University Press.

**Perrett G**. 1990. *The Language Testing Interview: A Reappraisal*. In De Jong J H A L & Stevenson D K (Eds). Individualising the Assessment of Language Abilities. Multilingual Matters. 225-238.

**Pollard J D E**. 1994. *Paper on Proficiency Testing*. Testing Special Interest Group Newsletter. International Association of Teachers of English as a Foreign Language (IATEFL).

**Ross S and Berwick R**, 1990. *The discourse of accommodation in oral proficiency interviews*. Studies in Second Language Acquisition (SSLA) 14:159-76.
**Ross S**, 1992. *Accommodative questions in oral proficiency interviews*. Language Testing 9/2: 173-186.
**Selinker L**, 1972. *Interlanguage*. International Review of Applied Linguistics 10: 209-231.
**Schegloff E A**, 1995. *Discourse as an Interactional Achievement III: The Omnirelevance of Action*. In Jacoby S & Ochs E (Eds): *Special issue on co-construction*. Research on Language and Social Interaction 28/3:185-211.
**Spolsky B**, 1995. *Measured Words*. Oxford University Press.
**Underhill N**, 1987. *Testing Spoken Language*. Cambridge University Press.
**Van Lier L**, 1989. *Reeling, Writhing, Drawling, Stretching, and Fainting in Coils: Oral Proficiency Interviews as Conversation*. TESOL Quarterly, 23/3.
**Woken M & Swales J M**, 1989. *Expertise and Authority in Native-nonnative Conversations: The Need for a Variable Account*. In Gass S, Madden C, Preston D & Selinker L (Eds): *Variation in Second Language Acquisition: Volume 1 - Discourse and Pragmatics*. Multilingual Matters: 211-227.
**Young R**, 1995. *Conversational Styles in Language Proficiency Interviews*. Language Learning 45/1:3-42.
**Young R and Milanovic M**, 1992. *Discourse variation in oral proficiency interviews*. Studies in Second Language Acquisition (SSLA) 14:403-24.
**Zuengler J**, 1993. *Encouraging learners' conversational participation: the effect of content knowledge*. Language Learning 43: 403-32.